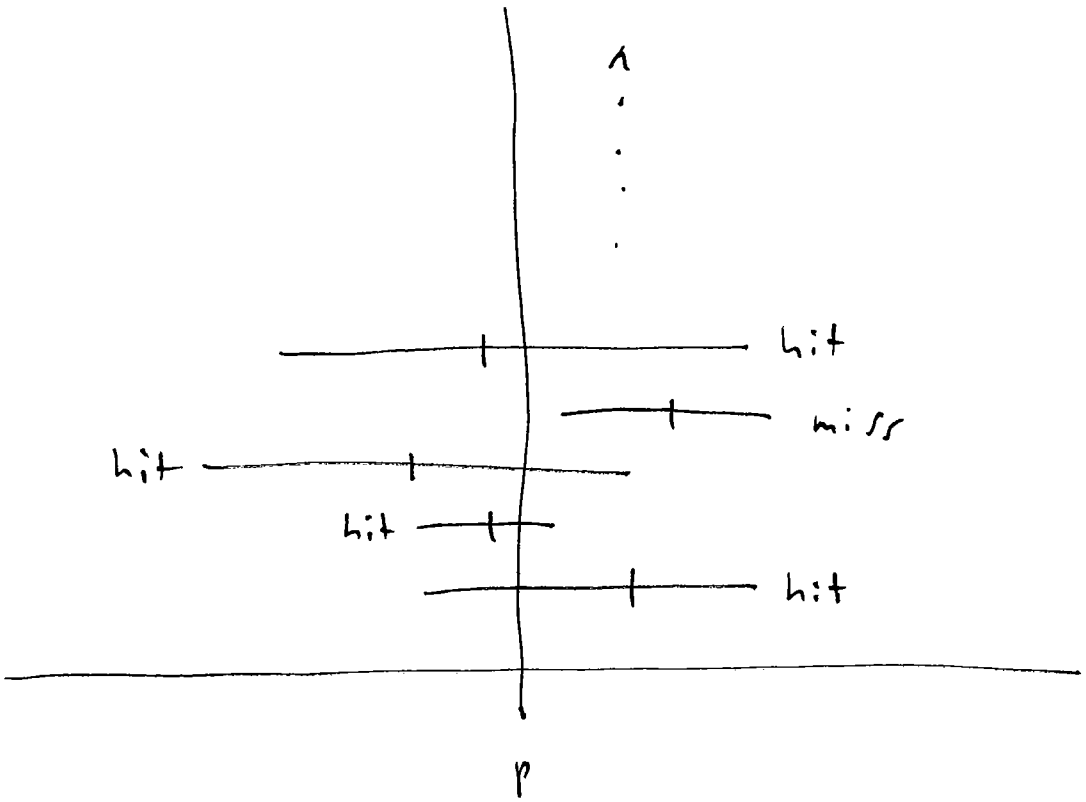
Frequentist Modeling (continued)

You would probably **want** the confidence interval (CI) to mean

$$P_F(0.142 \leq p \leq 0.218) \doteq 0.95, \qquad (6)$$

but it **can't** mean that in the frequentist approach to probability: in that approach p is treated as a **fixed unknown constant**, which either **is** or **is not** between 0.142 and 0.218.

So what **does** it mean?



This is a kind of **calibration** of the CI process: about 95% of the nominal 95% CIs would include the true value, if you were to generate a lot of them via independent IID samples from the population.

Frequentist Modeling (continued)

The diagram on page 6 takes up a lot of space; it would be nice to have a more **succinct summary** of it.

A random variable Y is said to follow the **Bernoulli distribution** with **parameter** $0 < p < 1$—this is summarized by saying $\boxed{Y \sim \mathrm{B}(p)}$—if Y takes on only the values 1 and 0 and

$$P(Y = y) = \left\{ \begin{array}{ll} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{array} \right\} = p^y (1 - p)^{1-y}. \qquad (7)$$

A **parameter** is just a **fixed unknown constant**.

Another **popular name** for the parameter p in this model is $\boxed{\theta}$.

Evidently what the population and sample parts of the diagram on page 6 are trying to say, in this notation, is that $(Y_1, \ldots, Y_n)$ are drawn **in an IID fashion** from the Bernoulli distribution with parameter θ.

In the usual **shorthand**, which I'll use from now on, this is expressed as

$$Y_i \overset{\text{IID}}{\sim} \mathrm{B}(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \qquad (8)$$

This is the **frequentist statistical model** for the AMI mortality data, except that we have forgotten so far to specify an important ingredient: **what is the population** of patients whose mean (underlying death rate) is θ?

As a frequentist (recall page 5), to use probability to quantify your uncertainty about the 1s and 0s, you have to think of them as either literally a **random sample** or **like** a random sample from some population, either hypothetical or actual.

Frequentist Modeling (continued)

What are some **possibilities** for this population?

- All AMI patients who **might have** come to the DH in 2000–03 if the world had turned out differently; or

- Assuming sufficient **time-homogeneity** in all relevant factors, you could try to argue that the collection of all 400 AMI patients at the DH from 2000–03 is **like** a random sample of size 400 from the population of all AMI patients at the DH from (say) 1997–2006; or

- **Cluster sampling** is a way to choose, e.g., patients by taking a random sample of hospitals and then a random sample of patients **nested** within those hospitals. What we actually have here is a kind of cluster sample of **all** 400 AMI patients from the DH in 2000–2003. Cluster samples tend to be less informative than SRS samples of the same size because of (positive) **intracluster correlation** (patients in a given hospital tend to be more similar in their outcomes than would an SRS of the same size from the population of all the patients in all the hospitals). Assuming the DH to be representative of some broader collection of hospitals in California and (unwisely) ignoring intracluster correlation, you could try to argue that these 400 1s and 0s were **like** a simple random sample of 400 AMI patients from this larger collection of hospitals.

None of these options is entirely **compelling**.

If you're willing to pretend the data are like a sample from some population, interest would then focus on inference about the **parameter** θ, the "underlying death rate" in this larger collection of patients to which you feel comfortable generalizing the 400 1s and 0s: if θ were unusually high, that would be **prima facie** evidence of a possible quality of care problem.

The Likelihood Function

Suppose (**as above**) that

$$Y_i \overset{\text{IID}}{\sim} \text{B}(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \tag{9}$$

Since the Y_i are **independent**, the **joint** sampling distribution of all of them, $P(Y_1 = y_1, \ldots, Y_n = y_n)$, is the **product** of the separate, or **marginal**, sampling distributions $P(Y_1 = y_1), \ldots, P(Y_n = y_n)$:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= P(Y_1 = y_1) \cdots P(Y_n = y_n) \\ &= \prod_{i=1}^{n} P(Y_i = y_i). \end{aligned} \tag{10}$$

But since the Y_i are also **identically distributed**, and each one is Bernoulli(θ), i.e., $P(Y_i = y_i) = \theta^{y_i}(1-\theta)^{1-y_i}$, the joint sampling distribution can be written

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}. \tag{11}$$

Let's use the symbol y to stand for the vector of **observed data values** $(y_1, \ldots, y_n)$.

Before any data have arrived, this joint sampling distribution is a function of y for fixed θ—it tells you **how the data would be likely to behave** in the future if you were to take an IID sample from the Bernoulli(θ) distribution.

The Likelihood Function (continued)

In 1921 Fisher had the following idea: **after** the data have arrived it makes more sense to interpret (11) as a function of θ for fixed y—he called this the **likelihood function** for θ in the Bernoulli(θ) model:

$$\begin{aligned} l(\theta|y) &= l(\theta|y_1,\ldots,y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \qquad (12) \\ &= P(Y_1 = y_1,\ldots,Y_n = y_n) \text{ but } \textbf{interpreted} \\ &\quad \textbf{as a function of } \theta \textbf{ for fixed } y. \end{aligned}$$

Fisher tried to create a theory of **inference** about θ **based only on this function**—we will see below that this is an important ingredient, **but not the only important ingredient**, in inference from the Bayesian viewpoint.

The Bernoulli(θ) likelihood function can be **simplified** as follows:

$$l(\theta|y) = \theta^s(1-\theta)^{n-s}, \qquad (13)$$

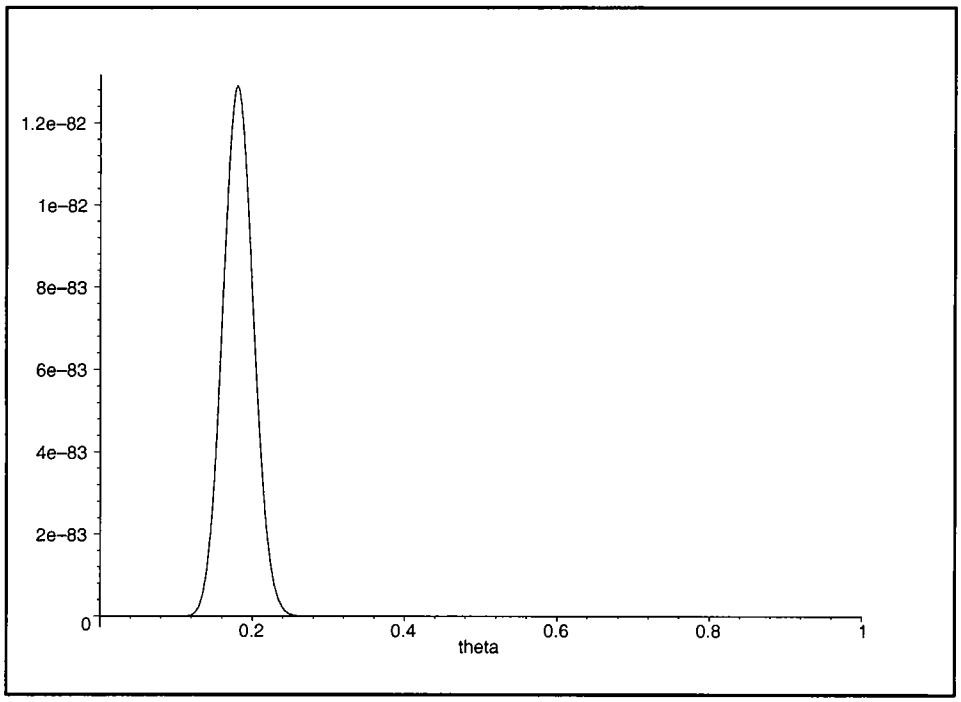
where $s = \sum_{i=1}^{n} y_i$ is the **number of 1s** in the sample and $(n-s)$ is the **number of 0s**.

What does this function **look like**?

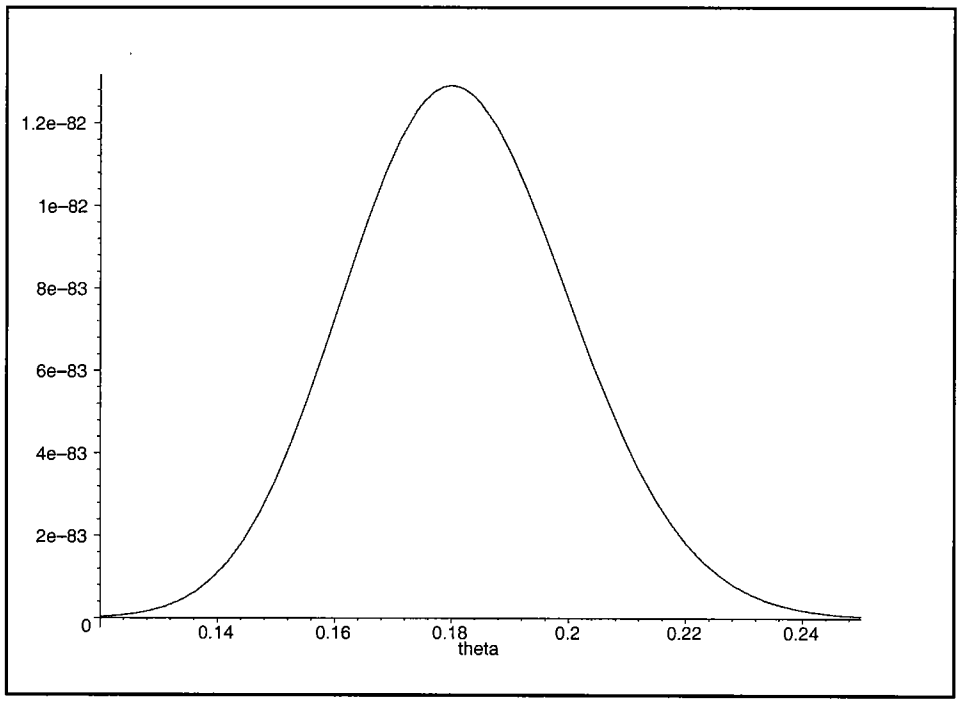
With $n = 400$ and $s = 72$ it's easy to get Maple to **plot it**:

```
rosalind 329> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> l := ( theta, s, n ) -> theta^s * ( 1 - theta )^( n - s );

                                        s            (n - s)
          l := (theta, s, n) -> theta  (1 - theta)

> plotsetup( x11 );

> plot( l( theta, 72, 400 ), theta = 0 .. 1 );
```

The Likelihood Function (continued)




```
> plot( l( theta, 72, 400 ), theta = 0.12 .. 0.25 );
```

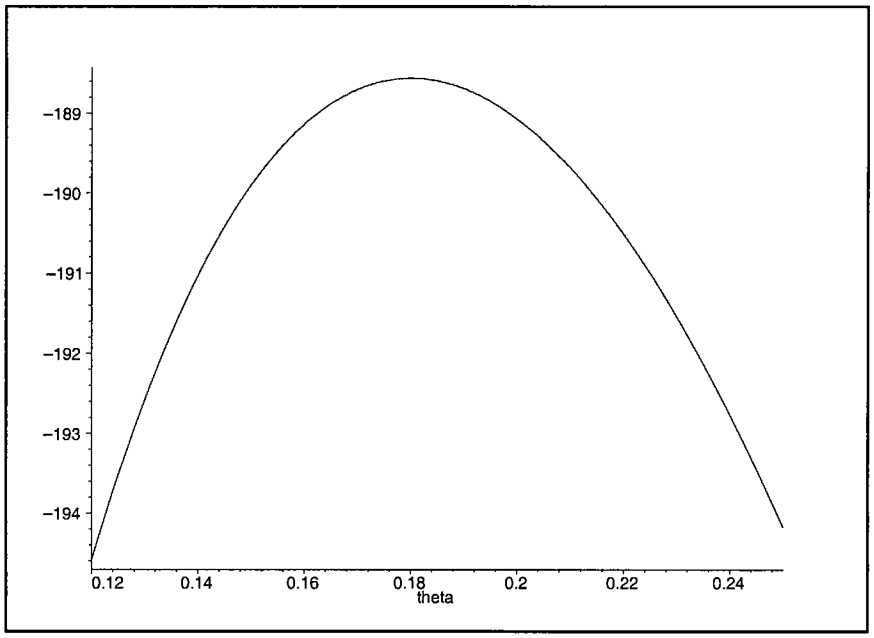


Does this function **remind you** of anything?

The Likelihood Function (continued)

It's often at least as useful to look at the **logarithm** of the likelihood function as the likelihood function itself:

```
> ll := ( theta, s, n ) -> log( l( theta, s, n ) );

> plot( ll( theta, 72, 400 ), theta = 0.12 .. 0.25 );
```



In this case, as is often true for large n, the log likelihood function looks **locally quadratic around its maximum**.

Fisher had the further idea that the **maximum** of the likelihood function would be a good **estimate** of θ (we'll look later at conditions under which this makes sense from the **Bayesian** viewpoint).

The Likelihood Function (continued)

Since the logarithm function is monotone increasing, it's equivalent in maximizing the likelihood to **maximize the log likelihood**, and for a function as well behaved as this you can do that by setting its first partial derivative with respect to θ to 0 and solving:

```
> score := simplify( diff( ll( theta, s, n ), theta ) );

                                  s - n theta
                    score := - ------------------
                               theta (-1 + theta)

> solve( score = 0, theta );

                                    s/n

> quit;

bytes used=2125632, alloc=1376004, time=0.51

rosalind 330>
```

The function of the data that maximizes the likelihood (or log likelihood) function is called the **maximum likelihood estimate** (MLE) $\widehat{\theta}_{\mathrm{MLE}}$.

Thus in this case $\widehat{\theta}_{\mathrm{MLE}}$ is just the **sample mean** $\frac{s}{n}$, which we've previously seen is a **sensible estimate** of θ.

Calibrating the MLE

Maximum likelihood provides a basic principle for estimation of a (population) parameter θ from the frequentist/likelihood point of view, but how should the **accuracy** of $\widehat{\theta}_{\text{MLE}}$ be assessed?

Evidently in the frequentist approach we want to compute the **variance** or **standard error** of $\widehat{\theta}_{\text{MLE}}$ in **repeated sampling**, or estimated versions of these quantities—let's focus on the estimated variance $\widehat{V}\left(\widehat{\theta}_{\text{MLE}}\right)$.

Fisher (1922) proposed an **approximation** to $\widehat{V}\left(\widehat{\theta}_{\text{MLE}}\right)$ that works well for large n and makes **good intuitive sense**.

In the AMI mortality case study, where $\widehat{\theta}_{\text{MLE}} = \widehat{\theta} = \frac{s}{n}$ (the **sample mean**), we already know that

$$V\left(\widehat{\theta}_{\text{MLE}}\right) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \widehat{V}\left(\widehat{\theta}_{\text{MLE}}\right) = \frac{\widehat{\theta}(1-\widehat{\theta})}{n}, \tag{14}$$

but Fisher wanted to derive results like this in a more **basic** and **general** way.

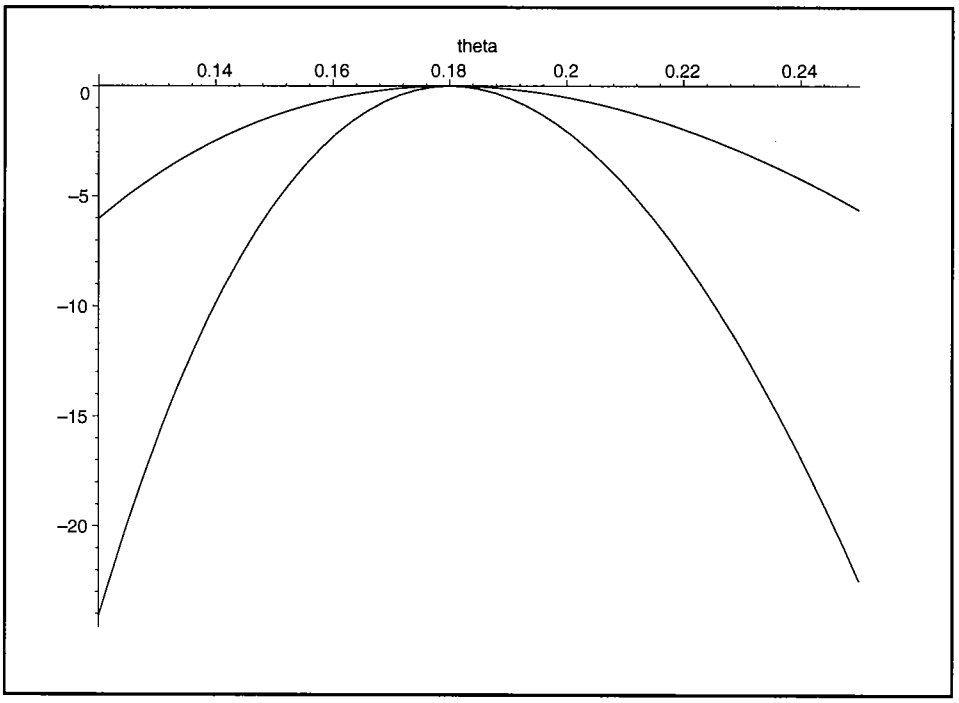
Calibrating the MLE (continued)

Imagine **quadrupling** the sample size in this case study from $n = 400$ to $n = 1600$ while keeping the observed death rate constant at 0.18—what would happen to the **log likelihood function**?

To answer this question, observe first that as far as maximizing the likelihood function is concerned it's equally good to work with **any (positive) constant multiple** of it, which is equivalent to saying that we can **add any constant** we want to the log likelihood function without harming anything.

In the Maple plot below I've added a **different constant** to each of the log likelihood functions with $(s, n) = (72, 400)$ and $(288, 1600)$ so that they both go through the point $(\widehat{\theta}_{\text{MLE}}, 0)$:

```
> plot( { ll( theta, 72, 400 ) - evalf( ll( 72 / 400, 72, 400 ) ),
    ll( theta, 288, 1600 ) - evalf( ll( 288 / 1600, 288, 1600 ) ) },
    theta = 0.12 .. 0.25, color = black );
```



Calibrating the MLE (continued)

Notice that what's happened as n went from 400 to 1600 while holding the MLE constant at 18% mortality is that the **second derivative of the log likelihood function at** $\hat{\theta}_{\text{MLE}}$ (a negative number) has **increased** in size.

This led Fisher to define a quantity he called the **information** in the sample about θ—in his honor we now call it the (observed) **Fisher information**:

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \left[-\frac{\partial^2}{\partial\theta^2}\log l(\theta|y)\right]_{\theta=\hat{\theta}_{\text{MLE}}}. \tag{15}$$

This quantity **increases** as n goes up, whereas our uncertainty about θ based on the sample, as measured by $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$, should go **down** with n.

Fisher conjectured and proved that the information and the estimated variance of the MLE in repeated sampling have the following simple **inverse relationship** when n is large:

$$\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) \doteq \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right). \tag{16}$$

He further proved that for large n (a) the MLE is approximately **unbiased**, meaning that in repeated sampling

$$E\left(\hat{\theta}_{\text{MLE}}\right) \doteq \theta, \tag{17}$$

and (b) the sampling distribution of the MLE is approximately **normal** with mean θ and estimated variance given by (16):

$$\hat{\theta}_{\text{MLE}} \dot{\sim} N\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right]. \tag{18}$$

Thus for large n an **approximate 95% confidence interval** for θ is given by $\hat{\theta}_{\text{MLE}} \pm 1.96\sqrt{\hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)}$.

Calibrating the MLE (continued)

You can **differentiate** to compute the information yourself in the AMI mortality case study, or you can use Maple to do it for you:

```
> score := ( theta, s, n ) -> simplify( diff( ll( theta, s, n ), theta ) );

      score := (theta, s, n) -> simplify(diff(ll(theta, s, n), theta))

> score( theta, s, n );

                                  s - n theta
                              - ------------------
                                theta (-1 + theta)

> diff2 := ( theta, s, n ) -> simplify( diff( score( theta, s, n ),
    theta ) );

    diff2 := (theta, s, n) -> simplify(diff(score(theta, s, n), theta))

> diff2( theta, s, n );

                                    2
                           -n theta  - s + 2 s theta
                           -------------------------
                                 2             2
                             theta  (-1 + theta)

> information := ( s, n ) -> simplify( eval( - diff2( theta, s, n ),
    theta = s / n ) );

> information( s, n );

                                      3
                                     n
                                - ----------
                                  s (-n + s)

> variance := ( s, n ) -> 1 / information( s, n );

                                                1
                      variance := (s, n) -> ----------------
                                            information(s, n)
```

Calibrating the MLE (continued)

```
> variance( s, n );

                      s (-n + s)
                    - ----------
                            3
                           n
```

This expression can be **further simplified** to yield

$$\widehat{V}\left(\widehat{\theta}_{\text{MLE}}\right) \doteq \frac{\frac{s}{n}\left(1-\frac{s}{n}\right)}{n} = \frac{\widehat{\theta}(1-\widehat{\theta})}{n}, \tag{19}$$

which **coincides** with (14).

From (19) **another expression** for the Fisher information in this problem is

$$\widehat{I}\left(\widehat{\theta}_{\text{MLE}}\right) = \frac{n}{\widehat{\theta}(1-\widehat{\theta})}. \tag{20}$$

As n increases, $\widehat{\theta}(1-\widehat{\theta})$ will tend to the constant $\theta(1-\theta)$ (this is well-defined because we've assumed that $0 < \theta < 1$, because $\theta = 0$ and 1 are probabilistically uninteresting), which means that information about θ on the basis of $(y_1, \ldots, y_n)$ in the IID Bernoulli model **increases at a rate proportional to n as the sample size grows**.

This is **generally true** of the MLE:

$$\widehat{I}\left(\widehat{\theta}_{\text{MLE}}\right) = O(n) \quad \text{and} \quad \widehat{V}\left(\widehat{\theta}_{\text{MLE}}\right) = O\left(n^{-1}\right), \tag{21}$$

as $n \to \infty$, where the notation $a_n = O(b_n)$ means that the ratio $\left|\frac{a_n}{b_n}\right|$ is bounded as n grows.

Thus uncertainty about θ on the basis of the MLE **goes down like $\frac{c_{\text{MLE}}}{n}$ on the variance scale** with more and more data (in fact Fisher showed that c_{MLE} achieves the lowest possible value: the MLE is **efficient**).

Bayesian Modeling

As a Bayesian in this situation, your job is to quantify your uncertainty about the 400 binary **observables** you'll get to see starting in 2000, i.e., your initial modeling task is **predictive** rather than inferential.

There is no samples-and-populations story in this approach, but probability and random variables arise in a different way: quantifying your uncertainty (for the purpose of betting with someone about some aspect of the 1s and 0s, say) requires **eliciting** from yourself a joint probability distribution that **accurately** captures your judgments about what you will see:

$$\boxed{P_{B:\text{you}}(Y_1 = y_1, \ldots, Y_n = y_n)}.$$

Notice that in the frequentist approach the random variables describe the **process** of observing a repeatable event (the "random sampling" appealed to here), whereas in the Bayesian approach you use random variables to quantify **your uncertainty about observables you haven't seen yet**.

I'll argue later that the concept of probabilistic **accuracy** has two components: you want your uncertainty assessments to be both **internally** and **externally** consistent, which corresponds to the Bayesian and frequentist ideas of **coherence** and **calibration**, respectively.

2.3 Exchangeability as a Bayesian concept parallel to frequentist independence

Eliciting a 400-dimensional distribution doesn't sound easy; major **simplification** is evidently needed.

In this case, and many others, this is provided by **exchangeability** considerations.

If (as in the frequentist approach) you have no relevant information that distinguishes one AMI patient from another, your uncertainty about the 400 1s and 0s is **symmetric**, in the sense that a random permutation of the **order** in which the 1s and 0s were labeled from 1 to 400 would leave your uncertainty about them unchanged. de Finetti (1930, 1964) called random variables with this property **exchangeable**:

$\{Y_i, i = 1, \ldots, n\}$ are **exchangeable** if the distributions of $(Y_1, \ldots, Y_n)$ and $(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ are the same for all permutations $(\pi(1), \ldots, \pi(n))$.

NB Exchangeability and IID are not the same: IID implies exchangeability, and exchangeable Y_i do have identical marginal distributions, but they're not independent (if you were expecting **a priori** about 15% 1s, say (that's the 30-day death rate for AMI with average-quality care), the knowledge that in the first 50 outcomes at the DH 20 of them were deaths would certainly change your prediction of the 51st).

de Finetti also defined **partial** or **conditional** exchangeability (e.g., Draper et al., 1993): if, e.g., the gender X of the AMI patients were available, and if there were evidence from the medical literature that 1s tended to be noticeably more likely for men than women, then you would probably want to assume **conditional** exchangeability of the Y_i given X (meaning that the male and female 1s and 0s, viewed as separate collections of random variables, are each unconditionally exchangeable). This is related to Fisher's (1956) idea of **recognizable subpopulations**.